INFORMATION CONDITIONING ADAPTATION BY MAXIMUM LIKELIHOOD MEAN LEARNING

Alexandru Murgu
Department of Mathematical Information Technology
University of Jyväskylä, FIN-40351 Jyväskylä, FINLAND

Abstract

In this paper we develop a mathematical statistics model for the mean learning processes occurring in adaptive systems. A maximum likelihood approach with penalty constraints and averaging is considered for increasing the performance in coping with the environment's uncertainties since it allows greater robustness by not relying on any particular prior assumption. The mean learning or averaging can be considered as a form of regularization since the effect of over-exposing the system to new information is reduced by averaging the predictions obtained from models which describe the local relevant information. Covariance evolution for dynamic systems having Markov parameters ensures that the data smoothing and information conditioning is encapsulated into the mean learning scheme.

Keywords: Expectation Maximization, Maximum Likelihood, Information Conditioning, Covariance Dynamics, Mean Learning, Adaptation, Prediction.

1. Introduction

The averaging ensembles of estimators to probability density estimation for Gaussian mixture models important methods in many learning and control applications operating in uncertain environments. The performance of averaging approach is enhanced by using the traditional regularization strategies such as the maximum likelihood approach and the Bayesian approach. In the maximum likelihood approach, some penalty functions can be derived using the conjugate Bayesian priors allowing the construction of the expectation maximization (EM) algorithms which can be used for learning purposes. The maximum likelihood approach with penalty constraints and the averaging increase the performance considerably compared to a standard maximum likelihood approach. The averaging is a superior way of coping with the environment's uncertainties since it allows greater robustness by not relying on any particular prior assumption. The averaging can be considered as a form of regularization since the effect of over-exposing the system to new information is reduced by averaging the predictions obtained from models which describe the local information exploration (Ormoneit and Tresp, 1998). The regularization is achieved by adding a penalty term to the log-likelihood cost function. In the Bayesian approach, the predictive distribution is approximated by averaging the forecasts of a sequence of parameter vectors selected according to the posterior probability density of the parameter vectors (White, 1996). In this respect, the Bayesian approach is related to both the regularization strategy (via the prior) and the averaging strategy (via models with different parameters). A covariance description of dynamic systems having Markov parameters ensures that the data smoothing and information conditioning will be exploited subsequently in order to match a prescribed model using an equivalent reduced dynamic system subject to a mean learning scheme (that is, averaging) using the EM solution to the smoothed maximum likelihood problem. The relevant local information is subsequently used in a mixture of normal multivariate distributions.

2. Information Conditioning for Data Smoothing

We consider a general formulation of the Expectation Maximization (EM) problem with multiplicative regularization (smoothing) of a data set. When the model data matrix is of maximum rank, the EM model has a data smoothing interpretation, that is, the fixed points of an iteration mapping (called the EMS solutions), solve a nonlinear system whose data have undergone a componentwise nonlinear smoothing. Given a large non-negative linear systems of the form

$$\mathbf{P}^T \underset{\sim}{\theta} = \mathbf{n}^* \tag{1}$$

with the model matrix $\mathbf{P} \in R^{B\times D}$, data vector $\mathbf{n}^* \in R^D$ and the solution $\underset{\sim}{\theta} \in R^B$ (all required to be non-negative), an arbitrary non-negative data smoothing system

$$\mathbf{Q}^T \underset{\sim}{\lambda} = \mathbf{n}^* \tag{2}$$

can be reformulated as in (1) using the following transformations

$$p_{bd} = \frac{q_{bd}}{r_b(\mathbf{Q})} \tag{3}$$

$$\theta_b = r_b(\mathbf{Q})\lambda_b \tag{4}$$

$$r_b(\mathbf{Q}) = \sum_d q_{bd} \tag{5}$$

These transformations allows the new coefficient matrix **P** to be normalized and to become row stochastic. We assume that **P** in equation (1) is row stochastic. In the context of probabilistic models, the EM algorithm is an iterative procedure whose iterates converge to a non-negative approximation of a solution to (1), independent of whether this system is over-, fully- or underdetermined. This algorithm generates a sequence of non-negative approximations to (1) as follows

$$\underset{\sim}{\theta}^{(n+1)} = \mathbf{F}(\underset{\sim}{\theta}^{(n)})\underset{\sim}{\theta}^{(n)}, \quad n = 0,\ 1,\ 2,\ldots \tag{6}$$

for a suitable initial selection $\underset{\sim}{\theta}^{(0)}$, where

$$\mathbf{F}(\underset{\sim}{\theta}) = \mathrm{diag}\left[F_1(\underset{\sim}{\theta}),\ldots,F_B(\underset{\sim}{\theta})\right] \tag{7}$$

and

$$F_b(\underset{\sim}{\theta}) = \sum_{d=1}^{D} \frac{n_d^* p_{bd}}{(\mathbf{P}^T\underset{\sim}{\theta})_d}\ , \quad b = 1,\ldots,B \tag{8}$$

The EM algorithm allows to find a non-negative solution called the EM solution of the *Maximum Likelihood Equations* (MLE)

$$\underset{\sim}{\theta} = \mathbf{F}(\underset{\sim}{\theta})\underset{\sim}{\theta} \tag{9}$$

The iteration (6) is noisy and slowly converging. An intermediate information conditioning step is introduced in order to obtain an EM with smoothing (EMS) algorithm. For *linear smoothing*, this algorithm takes the form

$$\underset{\sim}{\theta}^{(n+1)} = \mathbf{SF}(\underset{\sim}{\theta}^{(n)})\underset{\sim}{\theta}^{(n)}\ , \quad n = 0,\ 1,\ 2,\ldots \tag{10}$$

where $\mathbf{S} \in R^{B\times B}$ is a non-negative *conditioning matrix*. The fixed point iteration generates a sequence of non-negative iterates approximating a non-negative solution of the functional equation

$$\mathbf{SF}(\underset{\sim}{\theta})\underset{\sim}{\theta} = \underset{\sim}{\theta} \tag{11}$$

Remark 1. If the smoothing matrix $\mathbf{S}$ is chosen to be non-negative, row stochastic and irreducible, then we obtain the class of the *statistical smoothers* which, as a minimum are irreducible.

■

If $\mathbf{S}$ is invertible, every EMS solution $\underset{\sim}{\theta}^S$ lies in the convex region

$$\Omega_S = \left\{ \underset{\sim}{\theta} \geq \mathbf{0} \middle| \sum_b s_b \theta_b = N \right\} \tag{12}$$

where

$$N = \sum_d n_d^* \tag{13}$$

and

$$s_b = \sum_{b'} S_{b'b}^{-1} \tag{14}$$

Consequently, the set Ω_S, defining the constraints on the parameter $\underset{\sim}{\theta}$ as part of a hyperplane in R^B which is compact if all the quantities s_b are positive. We denote the *nonlinear EMS mapping* by

$$\underset{\sim}{\theta} \mapsto \mathbf{F}_S(\underset{\sim}{\theta}) \tag{15}$$

where

$$\mathbf{F}_S(\underset{\sim}{\theta}) = \mathbf{SF}(\underset{\sim}{\theta})\underset{\sim}{\theta} \tag{16}$$

We can write the system (11) in the form

$$\underset{\sim}{\theta} = \mathbf{S}\underset{\sim}{\Theta}\mathbf{P}\underset{\sim}{\lambda}(\underset{\sim}{\theta}) \tag{17}$$

where $\underset{\sim}{\Theta} = \operatorname{diag}\underset{\sim}{\theta}$. The vector $\underset{\sim}{\lambda} \in R^D$ is defined as

$$\lambda_d(\underset{\sim}{\theta}) = \frac{n_d^*}{(\mathbf{P}^T\underset{\sim}{\theta})_d} \quad , \quad d = 1, \ldots, D \tag{18}$$

The alternative way of representing the EMS iteration equation is

$$\underset{\sim}{\theta}^{(n+1)} = \mathbf{S}\underset{\sim}{\Theta}^{(n)}\mathbf{P}\underset{\sim}{\lambda}(\underset{\sim}{\theta}^{(n)}), \quad n = 0,\ 1, \ldots \tag{19}$$

Theorem 1. Assume that $\mathbf{P}$ has rank D and that $\mathbf{S}$ is non-negative and invertible. Then every positive EMS solution $\underset{\sim}{\theta}^S$ solves the nonlinear equation

$$\mathbf{P}^T\underset{\sim}{\theta} = \mathbf{n}(\underset{\sim}{\theta}) \tag{20}$$

where $\mathbf{n}(\underset{\sim}{\theta}) = \left[n_1(\underset{\sim}{\theta}), \ldots, n_D(\underset{\sim}{\theta}) \right]^T$ with

$$n_d(\underset{\sim}{\theta}) = \frac{n_d^*}{(\mathbf{P}^+(\mathbf{S}\underset{\sim}{\Theta})^{-1}\underset{\sim}{\theta})_d} \quad , \quad d = 1, \ldots, D \tag{21}$$

Every solution $\hat{\underset{\sim}{\theta}}$ of (20) without zero components solves the *modified* EMS *equations*

$$\mathbf{S}(\mathbf{F}(\underset{\sim}{\theta}) + \underset{\sim}{\Theta}^*)\underset{\sim}{\theta} = \underset{\sim}{\theta} \tag{22}$$

where $\Theta^* = \operatorname{diag}\underset{\sim}{\theta}^*$ for some $\underset{\sim}{\theta}^* \in \ker(\mathbf{P}^T)$.

■

Remark 2. Theorem 1 also holds for any *nonlinear* mapping $\mathbf{S}$ which is non-negative and locally invertible near θ^S, when $(\mathbf{S}\Theta^S)^{-1}$ is interpreted locally as being the composition $(\Theta^S)^{-1} \circ \mathbf{S}^{-1}$. If $\mathbf{S}$ is linear, the assumption $\theta^S > \mathbf{0}$ holds if, in addition to the given assumptions, $\mathbf{S}$ is irreducible.

■

The introduction of the nonlinearity $\mathbf{S}$ allows the EMS algorithm to perform very well for a suitable choice of $\mathbf{S}$. This situation resembles the phenomenon of super-resolution of near black objects by the *Maximum Entropy Method* (MEM). The intuition behind the data smoothing is the following: if $(\mathbf{S}\Theta^S)^{-1}\theta^S \approx \mathbf{1}_B$, then $\mathbf{P}^+(\mathbf{S}\Theta^S)^{-1}\theta^S \approx \mathbf{1}_D$ and so, $\mathbf{n}(\theta^S) \approx \mathbf{n}^*$. Any $\mathbf{S}$ which makes $(\mathbf{S}\Theta^S)^{-1}\theta^S - \mathbf{1}_B$ suitably small causes a small change in the data when the smoothing is performed to obtain $\mathbf{n}(\theta^S)$. The quantity $\left\|(\mathbf{S}\Theta^S)^{-1}\theta^S - \mathbf{1}_B\right\|$ is a measure of the amount of data smoothing.

Theorem 2. Assume that $\mathbf{S}$ is non-negative and invertible and N, $\mathbf{P}$ and

$$s_b = \sum_b S^{-1}_{bb}, \quad b = 1, \ldots, B \tag{23}$$

are all positive. Then the EMS equations have a solution $\theta^S \in \Omega_S$.

■

Theorem 3. The variation of any smooth EMS solution θ^S with the parameters $\mathbf{n}^*$ and $\mathbf{P}$ is the solution of the following systems

$$\mathbf{S}(\mathbf{I} - DF_S(\theta^S))\nabla_d\theta^S = \frac{1}{r^d}\Theta^S\mathbf{p}_d \quad , \quad d = 1, \ldots, D \tag{24}$$

$$\mathbf{S}(\mathbf{I} - DF_S(\theta^S))\nabla_{bd}\theta^S = \frac{1}{r^d}\Theta^S(\delta_b - \theta^S_b\mathbf{p}_d)n^*_d \quad , \quad b = 1, \ldots, B, \; d = 1, \ldots, D \tag{25}$$

where

$$\nabla_d\theta^S = \frac{\partial\theta^S}{\partial n^*_d} \tag{26}$$

$$\nabla_{bd}\theta^S = \frac{\partial\theta^S}{\partial p_{bd}} \tag{27}$$

$$(\delta_b)_{b'} = \delta_{bb'} \tag{28}$$

$$(\mathbf{p}_d)_{b'} = p_{b'd} \tag{29}$$

$$r^d = (\mathbf{P}^T\theta^S)_d \tag{30}$$

If $\mathbf{S}$ is chosen to be non-negative and invertible matrix such that $1 \notin \sigma(DF_S(\theta^S))$, then the systems (24)-(25) are uniquely solvable with respect to $\nabla_d\theta^S$ and $\nabla_{bd}\theta^S$.

■

A covariance description of dynamic systems having Markov parameters ensuring that the data smoothing and information conditioning is exploited in order to match a prescribed model is considered in the next section. An equivalent reduced dynamic system subject to a mean learning scheme (averaging) uses the solution of maximum likelihood problem.

3. Covariance Dynamics of Markov Systems

Truncation technique in dynamic systems based on observability matrices of the full order system are used to determine the order of reduced order model needed to match a specified number of output covariance derivatives and Markov parameters. The resulting realization is independent of the basis of the complete model up to a unitary transformation. Consider an asymptotically stable, controllable and observable stochastic system (Grimble and Johnson, 1988)

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{w}(t) \tag{31}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) \tag{32}$$

where $\mathbf{x} \in R^n$ is the state vector, $\mathbf{y} \in R^D$ is the measurement vector and $\mathbf{w} \in R^m$ is the noise vector. The parameter matrices **A**, **B**, **C** are constant and we assume that $\text{rank}[\mathbf{C}] = n_1 \le D$. The initial condition $\mathbf{x}(0)$ is assumed to be a zero mean Gaussian random vector with covariance matrix $\mathbf{X}_0$. The zero mean white noise process $\mathbf{w}(t)$ has the intensity $\mathbf{W} > \mathbf{0}$ and we assume that $\mathbf{w}(t)$ is independent of $\mathbf{x}(0)$. The covariance matrix of the output process $\mathbf{y}(t)$ in the presence of zero mean $\mathbf{x}(0)$, denoted by $\sum_{\tau 0}(t+\tau, t)$ is defined as

$$\sum_{\tau 0}(t,t) = E\left[\mathbf{y}(t)\mathbf{y}^T(t)\right] \tag{33}$$

and

$$\sum_{\tau 0}(t+\tau, t) = E\left[\mathbf{y}(t+\tau)\mathbf{y}^T(t)\right] \tag{34}$$

where $\sum_{\tau 0}(t,t)$ is calculated as

$$\sum_{\tau 0}(t,t) = \mathbf{C}\mathbf{X}(t,t)\mathbf{C}^T \tag{35}$$

with $\mathbf{X}(t,t)$, the covariance matrix of $\mathbf{x}(t)$ satisfying the following differential equation

$$\dot{\mathbf{X}}(t,t) = \mathbf{A}\mathbf{X}(t,t) + \mathbf{X}(t,t)\mathbf{A}^T + \mathbf{B}\mathbf{W}\mathbf{B}^T \tag{36}$$

with

$$\mathbf{X}(0,\ 0) = \mathbf{X}_0 \tag{37}$$

The *steady-state covariances* are defined as

$$\sum_{\tau 0} = \lim_{t\to\infty} \sum_{\tau 0}(t,t) \tag{38}$$

and

$$\mathbf{X} = \lim_{t\to\infty} \mathbf{X}(t,t) \tag{39}$$

where $\sum_{\tau 0}$ is given by

$$\sum_{\tau 0} = \mathbf{C}\mathbf{X}\mathbf{C}^T \tag{40}$$

and

$$\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^T + \mathbf{B}\mathbf{W}\mathbf{B}^T = \mathbf{0} \tag{41}$$

A *partial realization* of the system (31)-(32) is defined as

$$\dot{\mathbf{x}}_R(t) = \mathbf{A}_R\mathbf{x}_R(t) + \mathbf{B}_R\mathbf{w}(t) \tag{42}$$

$$\mathbf{y}_R(t) = \mathbf{C}_R\mathbf{x}_R(t) \tag{43}$$

where $\mathbf{x}_R \in R^r$, $\mathbf{y}_R \in R^D$ with $r \le n$ and

$$E\left[\mathbf{x}_R(0)\right] = \mathbf{0} \tag{44}$$

$$E\left[\mathbf{x}_R(0)\mathbf{x}_R^T(0)\right] = \mathbf{X}_{R0} \tag{45}$$

The system (42)-(43) is obtained by *truncating* (subsystem elimination) the system (31)-(32), that is, there exist the matrices $\mathbf{L}_R$ and $\mathbf{T}_R$ satisfying the following equations

$$\mathbf{L}_R\mathbf{T}_R = \mathbf{I}_r, \ \mathbf{L}_R \in R^{r\times n}, \ \mathbf{T}_R \in R^{n\times r} \tag{46}$$

such that

$$\mathbf{A}_R = \mathbf{L}_R\mathbf{A}\mathbf{T}_R \tag{47}$$

$$\mathbf{B}_R = \mathbf{L}_R\mathbf{B} \tag{48}$$

$$\mathbf{C}_R = \mathbf{C}\mathbf{T}_R \tag{49}$$

where $\mathbf{I}_r$ is the $(r\times r)$ identity matrix.

Definition 1. The realization (42)-(43) is called a q-COVariance Equivalent Realization (q-COVER) of (31)-(32) if and only if

$$E\left[\mathbf{y}_R(t)\right] = E\left[\mathbf{y}(t)\right] \tag{50}$$

and

$$\sum_{Rj} = \lim_{\tau\to 0}\left[\lim_{t\to\infty}\frac{d^j}{d\tau^j}\sum_{R0}(t+\tau,t)\right] = \lim_{\tau\to 0}\left[\lim_{t\to\infty}\frac{d^j}{d\tau^j}\sum_{0}(t+\tau,t)\right] = \sum_{j}, \ \ j=0,\ 1,\ldots,q \tag{51}$$

A minimal q-COVER is defined as satisfying (46)-(49) with the smallest possible order r.

■

We can write

$$\sum_{0}(t+\tau,t) = E\left[\mathbf{y}(t+\tau)\mathbf{y}^T(t)\right] = \mathbf{C}E\left[\mathbf{x}(t+\tau)\mathbf{x}^T(t)\right]\mathbf{C}^T = \mathbf{C}e^{\mathbf{A}\tau}\mathbf{X}(t,t)\mathbf{C}^T \tag{52}$$

and since $\mathbf{X}(t,t) = \mathbf{X}$ in the steady state, it follows that

$$\sum_{0}(\tau) = \lim_{t\to\infty}\sum_{0}(t+\tau,t) = \mathbf{C}e^{\mathbf{A}\tau}\mathbf{X}\mathbf{C}^T \tag{53}$$

Remark 3. Assuming the steady-state covariance conditions is *equivalent* to assuming that $\mathbf{y}(t)$ is a stationary process, since for stationary we have

$$E\left[\mathbf{y}(t+\tau)\mathbf{y}^T(t)\right] = \sum_{0}(t+\tau,t) = \sum_{0}(t+\tau-t) = \sum_{0}(\tau) \tag{54}$$

and

$$\sum_{0}(t,t) = \sum_{0} \tag{55}$$

■

From the equation (52) it follows that

$$\sum_{j}(\tau) = \lim_{t\to\infty}\frac{d^j}{d\tau^j}\sum_{0}(t+\tau,t) = \mathbf{C}e^{\mathbf{A}\tau}\mathbf{A}^j\mathbf{X}\mathbf{C}^T, \ \ j=0,\ 1,\ldots,q \tag{56}$$

$$\sum_{j} = \lim_{\tau\to 0}\sum_{j}(\tau) = \mathbf{C}\mathbf{A}^j\mathbf{X}\mathbf{C}^T, \ \ j=0,\ 1,\ldots,q \tag{57}$$

Theorem 4. A realization (42)-(43) is a q-COVER of (31)-(32) if and only if

$$\sum_{Rj} = \mathbf{C}_R\mathbf{A}_R^j\mathbf{X}_R\mathbf{C}_R^T = \mathbf{C}\mathbf{A}^j\mathbf{X}\mathbf{C}^T = \sum_{j}, \ \ j=0,\ 1,\ldots,q \tag{58}$$

■

Considering a series expansion of $e^{\mathbf{A}\tau}$ in (52), we can write

$$\lim_{t\to\infty}\sum_{\tau}(t+\tau,t)=\sum_{i=0}^{\infty}(\mathbf{C}\mathbf{A}^{i}\mathbf{X}\mathbf{C}^{T})\left(\frac{\tau^{i}}{i!}\right) \tag{59}$$

and similarly for the partial realization we get

$$\lim_{t\to\infty}\sum_{R0}(t+\tau,t)=\sum_{i=0}^{\infty}(\mathbf{C}_{R}\mathbf{A}_{R}^{i}\mathbf{X}_{R}\mathbf{C}_{R}^{T})\left(\frac{\tau^{i}}{i!}\right) \tag{60}$$

As q increases, the q-COVER forces more terms of $\sum_{R0}(t+\tau,t)$ to match those of $\sum_{0}(t+\tau,t)$ and (42)-(43) becomes a better partial realization of (31)-(32). The correlation over time of the reduced system outputs comes closer to matching the correlation over time of the full system outputs.

Remark 4. A q-COVER has the same input-output rate correlation as the full order system. This condition improves the fidelity of the reduced order model and it is regarded as an *internal smoothing mechanism* for model approximation. The steady state covariance of $\mathbf{x}_R(t)$ satisfies

$$\mathbf{X}_{R}\mathbf{A}_{R}^{T}+\mathbf{A}_{R}\mathbf{X}_{R}+\mathbf{B}_{R}\mathbf{W}\mathbf{B}_{R}^{T}=\mathbf{0} \tag{61}$$

■

Remark 5. The q-COVER is independent of the choice of basis of the state space of (31)-(32) up to a unitary transformation. Since a change of basis affects only the *endogenous variables* $\mathbf{x}(t)$, the q-COVERs depend only on the *exogenous variables* $\mathbf{y}(t)$ and $\mathbf{w}(t)$.

■

4. Maximum Likelihood Mean Learning

In the maximum likelihood approach we try to find the parameters which maximizes the function

$$l(\boldsymbol{\theta},\mathbf{y})=\log p(\boldsymbol{\theta},\mathbf{y})=\log p(\mathbf{y}|\boldsymbol{\theta})+\log p(\boldsymbol{\theta}) \tag{62}$$

with respect to $\boldsymbol{\theta}$, where the first term is the log-likelihood and the second is the logarithm of the prior parameter distribution. As a density model for the data we choose the class of Gaussian mixtures

$$p(\mathbf{y}|\boldsymbol{\theta})=\sum_{j=0}^{q}\alpha_{j}N(\mathbf{y}|\boldsymbol{\mu}_{j},\boldsymbol{\Sigma}_{j}) \tag{63}$$

with the restrictions $\alpha_i \geq 0$ and

$$\sum_{j=1}^{q}\alpha_{j}=1 \tag{64}$$

where $N(\mathbf{y}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$ are the multivariate normal densities

$$N(\mathbf{y}|\boldsymbol{\mu}_{j},\boldsymbol{\Sigma}_{j})=(2\pi)^{-\frac{d}{2}}\left|\boldsymbol{\Sigma}_{j}\right|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{j})^{T}\boldsymbol{\Sigma}_{j}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{j})\right] \tag{65}$$

The Gaussian mixture model is well suited to approximate a wide class of continuous probability densities (White, 1996). Given the data set $Y = \{\mathbf{y}_1,\ldots,\mathbf{y}_m\}$ of realizations for $\mathbf{y}$, we formulate the log-likelihood as

$$l(\boldsymbol{\theta},\mathbf{y}) = \log\left[\prod_{k=1}^{m} p(\mathbf{y}_k|\boldsymbol{\theta})\right] + \log p(\boldsymbol{\theta}) = \sum_{k=1}^{m}\log\sum_{j=0}^{q}\alpha_i N(\mathbf{y}_k|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) + \log p(\boldsymbol{\theta}) \tag{66}$$

Maximum likelihood parameter estimates $\boldsymbol{\theta}$ using the EM algorithm which consists of the iterative application of the following two steps:

1. **E-step.** Based on the current parameter estimates, the posterior probability that the component j of the covariance equivalent realization is responsible for the generation of data pattern $\mathbf{y}_k$ is estimated as

$$\beta_{jk} = \frac{\alpha_j N(\mathbf{y}_k|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}{\sum_{i=0}^{q}\alpha_i N(\mathbf{y}_k|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)} \tag{67}$$

2. **M-step.** The new parameter estimates are computed as

$$\alpha_j' = \frac{1}{m}\sum_{k=1}^{m}\beta_{jk} \tag{68}$$

$$\boldsymbol{\mu}_j' = \frac{\sum_{k=1}^{m}\beta_{jk}\mathbf{y}_k}{\sum_{k=1}^{m}\beta_{jk}} \tag{69}$$

$$\boldsymbol{\Sigma}_j' = \frac{\sum_{k=1}^{m}\beta_{jk}(\mathbf{y}_k-\boldsymbol{\mu}_j')(\mathbf{y}_k-\boldsymbol{\mu}_j')^T}{\sum_{k=1}^{m}\beta_{jk}} \tag{70}$$

Theorem 5. For a given i, the q-COVERs ($q > i$) satisfy the equation

$$\lim_{t\to\infty} E\left[\mathbf{y}_R^{(j)}(t)(\mathbf{y}_R^{(l)})^T\right] = \lim_{t\to\infty} E\left[\mathbf{y}^{(j)}(t)(\mathbf{y}^{(l)})^T\right], \quad j,\ l = 0,\ 1,\ldots,i \tag{71}$$

where

$$[\cdot]^{(j)} = \frac{d^j}{dt^j}[\cdot] \tag{72}$$

■

Remark 6. In the special case $q = 1$, the 1-COVER has the property of minimizing a quadratic criterion associated to the truncation error (when $n_1 = k$). Define the output error as

$$\mathbf{e}(t) = \mathbf{y}(t) - \mathbf{y}_R(t) \tag{73}$$

Assuming that $\mathbf{C}_R = \mathbf{I}_k$, the error $\mathbf{e}(t)$ satisfies the equation

$$\dot{\mathbf{e}}(t) = \mathbf{A}_R\mathbf{e}(t) + \mathbf{f}(t) \tag{74}$$

where

$$\mathbf{f}(t) = (\mathbf{CA} - \mathbf{A}_R\mathbf{C})\mathbf{x}(t) + (\mathbf{CB} - \mathbf{C}_R\mathbf{B}_R)\mathbf{w}(t) \tag{75}$$

The quadratic criterion

$$J = \min_{\mathbf{f}(t)} E\left[\int_0^t \mathbf{f}^T(t)\mathbf{f}(t)dt\right] \tag{76}$$

This minimization criterion can be regarded in two different ways as follows:

a) $\mathbf{f}(t)$ is the forcing term (learning stimulus) in the output residual error equation (65). The minimization is an appropriate goal which is associated to the learning/approximation of the full system model.

b) $\mathbf{f}(t)$ can be considered as the equation error resulting from substituting the output into the equation of the reduced model, that is,

$$\mathbf{f}(t) = \dot{\mathbf{y}}(t) - \left[\mathbf{A}_R\mathbf{y}(t) + \mathbf{B}_R\mathbf{w}(t)\right] \tag{77}$$

In this case, the minimization (67) becomes a linear problem called the output learning model (Ormoneit and Glynn, 2002).

■

In the case of impulse or white noise inputs, the parameter matrices $\mathbf{A}_R$, $\mathbf{B}_R$ which minimizes J are given by

$$\mathbf{A}_R = \mathbf{CAXC}^T(\mathbf{CXC}^T)^{-1} \tag{78}$$

and

$$\mathbf{B}_R = \mathbf{CB} \tag{79}$$

Remark 7. The truncation technique provides a link between the matching of Markov parameters to the output covariance derivatives in the partial realization problem. This technique a q-COVER which simultaneously matches $q+1$ output covariance derivatives and q Markov parameters.

■

The aim of the next section is to provide a basis for identifying the elements of an adaptation model based on observations from a dynamic system and statistical inference drawn from a postulated reference model. The issue is to have a robust inference from observations as a self-adaptation when learning the truly valuable new information contained in the data. As describe in the previous section, this issue is handled via the maximum likelihood averaging by detecting the system pattern behaviour. The observations data can contain non-typical records (i.e., outliers) which are regarded as being innovations in the learning process. These outliers' potential contribution to knowledge formation process is evaluated locally in conjunction with the latest estimation of the knowledge map prior to the sampling of the new observation.

The statistical inference drawn from the data can be substantially altered when the observation evaluated as the outlier is deleted from the knowledge map formation process. The outlier generation process could have a mean-shift model or a variance-inflation model. In other to deal with the forthcoming complexities in the data, we consider a general regression model as the basic hypothesis for the observation process. To assess the practical advantage of the averaging and regularization approaches, the density estimates of the multivariate statistical processes are used to obtain the control model for adaptation by learning. The reason is that the generalization error of density estimates in terms of the likelihood based on test data is rather unintuitive.

5. Application to System Adaptation by Learning

Let us consider an exponomial function approximation $\mathbf{p}(t)$ defined as

$$\mathbf{p}(t) = \sum_{j=1}^{q} c_j^t \mathbf{x}_j, \quad t = 1,\ 2, \ldots, m \tag{80}$$

where the c_j are fixed and distinct real numbers. Suppose that we have m observations on a nonstationary stochastic process $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m]^T$ with finite second order moments, where $\mathbf{y}_m$ is the first observation, $\mathbf{y}_{m-1}$ the second, etc., and $\mathbf{y}_1$ the last observation in the series. Assume that the series of observations obeys the model

$$\mathbf{y}_t = \mathbf{p}(t) + \mathbf{e}_t, \quad t = 1,\ 2, \ldots,\ m, \quad m > q \tag{81}$$

Let $\mathbf{C}$ be the generalized Vandermonde matrix

$$\mathbf{C} = \begin{pmatrix} c_1 & c_2 & \cdots & c_q \\ c_1^2 & c_2^2 & \cdots & c_q^2 \\ \vdots & \vdots & \cdots & \vdots \\ c_1^m & c_2^m & \cdots & c_q^m \end{pmatrix} \tag{82}$$

and let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q]^T$ and $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]^T$. The model can be reformulated into a matrix notation

$$\mathbf{Y} = \mathbf{C}\mathbf{X} + \mathbf{E} \tag{83}$$

where $E[\mathbf{E}] = \mathbf{0}$ and $\mathrm{Var}[\mathbf{E}] = E[\mathbf{E}\mathbf{E}^T] \equiv \sigma^2 \Sigma^{-1}$. The covariance matrix is assumed to be positive definite. The matrix $\mathbf{C}$ has full rank because the c_j are distinct. If Σ^{-1} is not the identity matrix, then $\mathbf{A}$ cannot be estimated simply by the usual least squares technique. For the estimates so obtained will *not* in general have minimum variance among all unbiased linear estimates. To obtain the best linear unbiased estimator (BLUE), a weighted least squares analysis is necessary. It can be shown that there exists a nonsingular symmetric matrix $\mathbf{G}$ such that

$$\mathbf{G}^T\mathbf{G} = \mathbf{G}\mathbf{G} = \Sigma \tag{84}$$

and multiplying $\mathbf{Y}$ by $\mathbf{G}$, we have

$$\mathbf{G}\mathbf{Y} = \mathbf{G}\mathbf{C}\mathbf{X} + \mathbf{G}\mathbf{E} \tag{85}$$

Since $E[\mathbf{G}\mathbf{E}] = \mathbf{0}$ and $\mathrm{Var}[\mathbf{G}\mathbf{E}] = E[\mathbf{G}\mathbf{E}\mathbf{E}^T\mathbf{G}^T] = \sigma^2\mathbf{I}$, $\mathbf{X}$ is found by minimizing

$$(\mathbf{G}\mathbf{E})^T\mathbf{G}\mathbf{E} = (\mathbf{Y} - \mathbf{C}\mathbf{X})^T\Sigma(\mathbf{Y} - \mathbf{C}\mathbf{X}) \tag{86}$$

with respect to $\mathbf{X}$. Then we have

$$\hat{\mathbf{X}} = (\mathbf{C}^T\Sigma\mathbf{C})^{-1}\mathbf{C}^T\Sigma\mathbf{Y} \tag{87}$$

where obviously, the inverse exists since $\mathbf{C}$ has full rank and Σ is positive definite. Define $\Sigma^{1/2}$ to be the positive definite square root of Σ and let $\mathbf{D} = \Sigma^{1/2}\mathbf{C}$. Then, we see at once also that

$$\hat{\mathbf{X}} = \mathbf{D}^{+}\Sigma^{1/2}\mathbf{y} \tag{88}$$

where $\mathbf{D}^+$ is the generalized inverse of $\mathbf{D}$ and consequently,

$$\hat{\mathbf{X}} = \mathbf{X} + (\mathbf{C}^T \underset{\sim}{\Sigma} \mathbf{C})^{-1} \mathbf{C}^T \underset{\sim}{\Sigma} \mathbf{E} \tag{88}$$

where $E\left[\hat{\mathbf{X}}\right] = \mathbf{X}$ and $\text{Var}(\hat{\mathbf{X}}) = \sigma^2 (\mathbf{C}^T \underset{\sim}{\Sigma} \mathbf{C})^{-1}$. Let $\hat{\mathbf{E}} = \mathbf{Y} - \mathbf{C}\hat{\mathbf{X}}$ denote the estimation error matrix. If $\mathbf{Y}$ has a multivariate normal distribution, then it can be shown that $\hat{\mathbf{X}}$ and $\hat{\mathbf{E}}^T \underset{\sim}{\Sigma} \hat{\mathbf{E}} / m$ are the maximum likelihood estimate of $\mathbf{X}$ and $\sigma^2 \mathbf{I}$. In this case, we can also perform the usual significance tests and construct confidence intervals for the $\mathbf{x}_j$ (White, 1996). We construct the one-step predictor for $\mathbf{y}_0$ with the purpose of on-line adaptation (Cassandras and Lafortune, 1999). The best (in the sense of minimum variance) linear unbiased one-step predictor of $\mathbf{y}_0$, given the previous observations $\mathbf{y}_1$, $\mathbf{y}_2$,…, $\mathbf{y}_m$. We assume that

$$\mathbf{y}_0 = \mathbf{1}^T \mathbf{X} + \mathbf{e}_0 \tag{89}$$

where $E\left[\mathbf{e}_0\right] = \mathbf{0}$, $\text{Var}\left[\mathbf{e}_0\right] = \sigma_0^2 \mathbf{I}$ and $\mathbf{1} = \left[1, 1, \ldots, 1\right]^T$ and define the matrix of expectations $\mathbf{V}$ by

$$\mathbf{V} = E\left[\mathbf{e}_0^T \mathbf{E}\right] \tag{90}$$

The best linear unbiased predictor of $\mathbf{y}_0$, denoted by $\hat{\mathbf{y}}_0$ is

$$\hat{\mathbf{y}}_0 = \mathbf{1}^T \hat{\mathbf{X}} + \frac{1}{\sigma^2} \mathbf{V}^T \underset{\sim}{\Sigma} \mathbf{V} \tag{91}$$

In the case when $\mathbf{e}_0$ is not correlated with $\mathbf{E}$, $\hat{\mathbf{y}}_0$ reduces to the one-step forecast, namely

$$\mathbf{y}_0^* = \mathbf{1}^T \hat{\mathbf{X}} \tag{92}$$

where

$$E\left[\mathbf{y}_0^*\right] = \mathbf{1}^T \mathbf{X} \tag{93}$$

$$\text{Var}\left[\mathbf{y}_0^*\right] = \sigma^2 \mathbf{1}^T (\mathbf{C}^T \underset{\sim}{\Sigma} \mathbf{C})^{-1} \mathbf{1} \tag{94}$$

In the same way we can construct the $(k+1)$th step forecast. If it is assumed that $\mathbf{y}_{-k} = \mathbf{c}^T \mathbf{X} + \mathbf{e}_{-k}$, where $\mathbf{c} = \left[c_1^{-k}, c_2^{-k}, \ldots, c_m^{-k}\right]^T$, $E\left[\mathbf{e}_{-k}\right] = \mathbf{0}$, $\text{Var}\left[\mathbf{e}_{-k}\right] = \sigma_{-k}^2 \mathbf{I}$, then

$$\hat{\mathbf{y}}_{-k} = \mathbf{c}^T \mathbf{X} + \sigma_{-k}^{-2} \underset{\sim}{\Omega}^T \underset{\sim}{\Sigma} \mathbf{E} \tag{95}$$

where $\underset{\sim}{\Omega} = E\left[\mathbf{e}_{-k}^T \mathbf{E}\right]$. The regression model of the environment is a way of encoding how the environment will interact with the adaptative/learning system. The time sampling strategy of the data produces the probabilistic description of the knowledge needed by the system in order to mimic or simulate the experience accumulation.

The mixture of normal multivariate distributions coupled with a regression estimate for the model parameters as in Section 4 is allows to perform mean learning in the presence of multiple operating modes describing the overall behaviour of the adaptive system. The one-step or multi-step ahead prediction framework remains in the general spirit of the parameter estimation theory for linear-quadratic systems (Harvey, 1996). The adaptation model of the closed loop for knowledge growth of the adaptive features is needed for the optimal evolution in an uncertain environment.

6. Concluding Remarks

This paper represents an attempt to close some earlier developments on the use of the maximum likelihood strategies for on-line information processing for control synthesis in systems operating in uncertain environments such as the telecommunication traffic flows (Murgu, 1995). The averaging technique previously considered in the context of ordered means for dynamics of normal populations (Murgu, 2001a) and the Kalman filtering for mixture systems (Murgu, 2001b) is regarded now as a data smoothing/regularization method for which the expectation maximization algorithms are available in the literature. The dynamic covariance modeling for equivalent Markov system realization is involved into the matrix statistical data smoothing of the EMS solutions. This type of data smoothing approach can be related to the maximum entropy method which is widely used as spectral optimization criteria for statistical control systems. Further research efforts along this path will lead to additional insights on the probabilistic nature of the learning models and knowledge formation from experimental data.

References

Cassandras C.G. and S. Lafortune (1999); Introduction to Discrete Event Systems; Kluwer Academic Publishers.

Grimble M.J. and M.A. Johnson (1988); Optimal Control and Stochastic Estimation. Theory and Applications, Vol. 1, John Wiley & Sons.

Harvey A.C. (1996); Forecasting, Structural Time Series Models and the Kalman Filter; Cambridge University Press.

Murgu A. (1995); "Maximum Likelihood Estimation from Multiple Observations for Neural Control of Traffic Flows in Multistage Networks"; Advances in Interdisciplinary Studies in Systems Research and Cybernetics, IIAS, G.E. Lasker (ed.), Vol. III, pp. 55-64.

Murgu A. (2001a); "Interval Estimation of Ordered Means fir Dynamics of Normal Populations"; Proceedings of the Focus Symposium on "Adaptive, Cooperative and Competitive Processes in System Modeling, Design and Analysis", A. Murgu, G.E. Lasker (eds.), IIAS, pp. 19-27.

Murgu A. (2001b); "Kalman Filtering of Mixture Systems with Interval Dynamics"; Proceedings of the Focus Symposium on "Adaptive, Cooperative and Competitive Processes in System Modeling, Design and Analysis", A. Murgu, G.E. Lasker (eds.), IIAS, pp. 28-36.

Ormoneit D. and V. Tresp (1998); "Averaging, Maximum Penalized Likelihood and Bayesian Estimation for Improving Gaussian Mixture Probability Density Estimates"; IEEE Trans. on Neural Networks, Vol. 9, No. 4, pp. 639-650.

Ormoneit D. and P. Glynn (2002); "Kernel-Based Reinforcement Learning in Average-Cost Problems"; IEEE Trans. on Automatic Control, Vol. 47, No. 10, pp. 1624-1636.

White H. (1996); Estimation, Inference and Specification Analysis; Cambridge University Press.